A Paradox in Decision-Theoretic Interval Estimation

BU-1086-M

March, 1989

Revised June, 1990

George Casella[1]

Jiunn Tzon Hwang[2]

Christian Robert[3]

Cornell University and Université Paris VI

Key Words and Phrases: Confidence Sets, Decision Theory, Bayes Estimation, Foundations

AMS 1980 Subject Classification: Primary 62C05, Secondary 62F25, 62A99

[1]Research supported by National Science Foundation Grant No. DMS89-0039

[2]Research supported by National Science Foundation Grant No. DMS88-09016

[3]Research supported by the U. S. Army Research Office through the Mathematical Sciences Institute at Cornell University

Summary

Decision-theoretic interval estimation has usually used a loss function that is a linear combination of volume and coverage probability. Such loss functions, however, may result in paradoxical behavior of the Bayes rules. We investigate this paradox in the case of Student's t, and suggest ways of avoiding it using a different loss function. Some properties of the resulting Bayes rules are also examined. This alternative approach may also be generalized.

1. Introduction

A set estimator for a parameter θ, based on observing X=x according to some distribution $f(x|\theta)$, is a set C_x in the parameter space Θ. The question of measuring optimality (either frequentist or Bayesian) of a set estimator against a loss criteria combining size and coverage does not yet have a satisfactory answer. For the case of Student's t interval, we examine some difficulties with the commonly used linear loss function and suggest alternative loss functions which eliminate these problems.

There are a number of advantages to a loss function approach to set estimation. From a theoretical view, this is the simplest way to address optimality properties such as admissibility or minimaxity. Furthermore, derivation of Bayes or generalized Bayes sets is straightforward (for example, as in Berger, 1980, Casella and Hwang, 1983, or Meeden and Vardeman, 1985). From a practical view, consideration of a meaningful loss function would allow interplay between the size and coverage components. This avoids possibly undesirable behavior that may occur if the components are considered separately. (For example, bounding coverage probability then optimizing size might lead to a set estimator whose size is too large to be of use.)

Although there has been much research into optimal set estimation, no satisfactory loss function has emerged. Most researchers who have combined size and coverage have used losses of the form

$$L(\theta,C) = a\cdot\mathrm{vol}(C) - I(\theta \in C) \quad , \quad a > 0, \tag{1.1}$$

where vol(C) denotes the volume of the set C,

$$I(\theta\in C) = \begin{cases} 1 & \text{if } \theta \in C \\ 0 & \text{if } \theta \notin C \end{cases},$$

and a is a fixed constant. Although a loss function of this type is reasonable to work with theoretically (see, for example, Joshi, 1969), and can sometimes be related to a componentwise loss (Casella and Hwang, 1982, Cohen and Sackrowitz, 1984), this loss can lead to a paradox, as shown in Section 2.

Fortunately, the more general class

$$L_S(\theta,C) = S\Big(\mathrm{vol}(C)\Big) - I(\theta \in C)\,, \tag{1.2}$$

where $S(\cdot)$ is an appropriately chosen nonlinear, nondecreasing, size function can eliminate the paradox. We give conditions on losses of the form (1.2) to obtain this more coherent behavior.

The history of optimal set estimation is long and varied, dating back (at least) to

Wilks (1938), who investigated optimal likelihood regions. Decision-theoretic treatments of the set estimation problem are contained in Blyth (1951), Brown (1966) and Joshi (1967, 1969). Blyth and Joshi were tangentially concerned with the relationship between the linear combination loss of (1.1), and a vector valued loss like

$$L_V(\theta,C) = \Big(\text{vol}(C), I(\theta \in C)\Big). \tag{1.3}$$

This relationship was further explored in Casella and Hwang (1982), where some correspondences between admissibility and minimaxity were derived. Also, Cohen and Sackrowitz (1984) established a relationship between $L_V(\theta,C)$ of (1.3) and another type of single-valued loss, one that introduces an auxiliary parameter. Other decision-theoretic approaches to set estimation, based on linear combination losses like (1.1), have been given by Winkler (1972), Cohen and Strawderman (1973b), and Meeden and Vardeman (1985).

In contrast to the loss function approach, other authors have worked directly with the vector loss (1.3). Most often, the technique is to restrict consideration to set estimators satisfying a minimum coverage probability requirement and, within this class, to optimize volume (or some other measure of size). These types of considerations also have a long history, being considered by Neyman (1937) and the aforementioned Wilks (1938). Different distributions have been considered by many authors, for example, Sterne (1954) looked at the binomial and Tate and Klett (1959) looked at a normal variance. Pratt (1961) showed, among other things, the relationship between volume and false coverage. The decision theoretic implications of this relationship was explored by Cohen and Strawderman (1973a), who showed that admissibility using the pair {probability of true coverage, volume} implies admissibility using the pair {probability of true coverage, probability of false coverage}. Then Stein (1962) explained how the usual confidence sphere for a multivariate normal mean could be dominated under the vector valued loss (1.3). This led to the papers of Brown(1966) and Joshi (1967) who established existence of dominating sets, and Hwang and Casella (1982, 1984), who exhibited such sets. Taking a slightly different approach, relying on invariance arguments, Hooper (1982) derived best invariant sets.

We might ask, at this point, what are the shortcomings of (1.1) and (1.3), and how can they be remedied by considering (1.2)? Although consideration of the individual loss components is very important, the vector-valued loss allows no interplay of volume and coverage, which makes it restrictive. Since one component must be fixed and the other optimized there is no jointly optimal solution. Also, consideration of the vector loss complicates the decision-theoretic comparisons, as the risks are no longer single-valued. (The vector loss function is not free of paradoxical behavior either. As shown by Casella and Hwang (1986) vector loss, along with many other losses, allows a type of paradox based on the Stein effect.)

The linear combination loss (1.1) eliminates the complications of a vector-valued loss, but introduces serious problems of its own, which will be addressed in the next section. It might also appear that the approach of Cohen and Sackrowitz (1984) would allow interplay. Their loss function, however, contains an unknown auxiliary parameter, so the actual value of the loss is not available to the experimenter. Their approach equates vector loss with a *class* of single-valued loss functions. Although consideration of a class of losses will implicitly involve the relationship between size and coverage, in actuality it still results in consideration of both components separately.

Loss functions of the form (1.2) can solve our problems, and allow the experimenter to describe the desired relationship between volume and coverage. Also, it is possible to put bounds on the ranges of the optimal sets (for a given S), bounds that will give either a minimal length or minimal coverage. Furthermore, adoption of the S function eliminates the undesirable behavior of the linear combination loss, and allows us to evaluate set estimators using a single-valued loss function.

In Section 2 we describe a paradox associated with the loss (1.1), and in Section 3 conditions are given, on a loss of the form (1.2), which can eliminate the paradox. Section 4 establishes a few results concerning the general behavior of the Bayes sets using losses of the form (1.2). A more thorough development of decision-theoretic properties under losses (1.2) can be found in Casella, Hwang and Robert(1990).

2. A Paradox Related to the Linear Loss Function

We now show the paradoxical behavior of the linear combination loss of (1.1) when estimating a univariate normal mean with unknown variance.

If $X_1,\cdots,X_n$ are iid $n(\mu,\sigma^2)$, the interval

$$C_t(\bar{x},s) = \left\{\mu : \bar{x} - t\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t\frac{s}{\sqrt{n}}\right\}, \tag{2.1}$$

where $\bar{x} = \Sigma x_i/n$ and $s = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{1/2}$, is a Bayes highest posterior density (HPD) region against the improper prior

$$\pi(\mu,\sigma^2) = \frac{1}{\sigma^2}\, d\mu d\sigma^2 .$$

The posterior distribution of $\sqrt{n}(\mu-\bar{x})/s$ is T_{n-1}, Student's t with n-1 degrees of freedom. Moreover, the frequentist (unconditional) distribution of $\sqrt{n}(\bar{x}-\mu)/s$ is also T_{n-1}. Thus, the frequentist and Bayesian answers agree at a widely respected statistical procedure.

Here now is a paradox (or, at the very least, an undesirable feature), first pointed out by J.O. Berger (personal communication). Consider the loss function of (1.1). For $C_t(\bar{x},s)$ of (2.1), the posterior expected loss is given by

$$L\left(\theta, C_t(\bar{x},s) \mid \bar{x},s\right) = a\left(\frac{2ts}{\sqrt{n}}\right) - P\left[\mu \in C_t(\bar{x},s) \mid \bar{x},s\right]. \tag{2.2}$$

The set estimator $C_t(\bar{x},s)$ can be uniformly dominated in posterior expected loss by the set estimator

$$C_t'(\bar{x},s) = \begin{cases} C_t(\bar{x},s) & \text{if} \quad s \le \frac{\sqrt{n}}{2ta} \\ \{\bar{x}\} & \text{if} \quad s > \frac{\sqrt{n}}{2ta} \end{cases} \tag{2.3}$$

But $C_t'(\bar{x},s)$ is a ridiculous estimator, which is even more apparent when we realize that $\{\bar{x}\}$ can be replaced by $\emptyset$ or $\{17\}$. If s becomes large, indicating uncertainty, $C_t'(\bar{x},s)$ indicates certainty in the estimation of μ by collapsing to a point. Clearly there is a problem with such an estimator, as increased uncertainty in the data should lead to increased uncertainty in the set estimator. Even though $\emptyset$ and the parameter space Θ $(=\Re)$ are formally equivalent answers with respect to the loss function (and are equally useless to the experimenter, as they are both "noninformative"), they are intuitively different. We think of $\emptyset$ as the limiting case of "precise" sets, and Θ as the limiting case of "imprecise" sets.

Not only does the estimator (2.3) dominate a Bayes HPD region, it dominates Student's t interval (in the Bayesian sense and, for $a > (1-\alpha)/E(2ts/\sqrt{n})$, in the frequentist sense). Thus, we have a case where a disconcerting rule dominates a time-honored statistical procedure. The only reasonable conclusion is that there is a problem with the loss function.

The Bayes rule associated with the loss (1.1) has the same disconcerting behavior. Minimization of the posterior expected loss (2.2) leads to the HPD region

$$C_{t^*}(\bar{x},s) = \left\{\mu : \bar{x} - t^*\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t^*\frac{s}{\sqrt{n}}\right\}, \tag{2.4}$$

where $t^* = t^*(s)$ is either the unique solution of

$$\frac{as}{\sqrt{n}} - f_{n-1}(t^*) = 0, \qquad 0 < t^* < \infty,$$

where $f_{n-1}(\cdot)$ denotes Student's t density with n-1 degrees of freedom, or $t^*(s) = 0$ if

$$0 < \inf_t\left[\frac{as}{\sqrt{n}} - f_{n-1}(t)\right] = \frac{as}{\sqrt{n}} - f_{n-1}(0).$$

Clearly $t^*(s)$ decreases as s increases and, moreover, is equal to zero (with positive probability) for s large enough. Thus, the Bayes set (2.4) exhibits behavior similar to $C_t'(\bar{x},s)$ in that its size decreases as uncertainty increases. Note further that the value of a really plays no role. As long as a is a fixed value the aberrant behavior persists. (Of course,

decision theory can accommodate a being a function of the data, but the experimenter should be able to examine the loss function, and consider its consequences, before the experiment is performed.)

Thus we have the paradox. Our intuition would lead us to use the t interval $C_t(\bar{x},s)$ but a formal, statistically sound, derivation leads us to a nonintuitive interval such as $C_t'(\bar{x},s)$. The obvious candidate for blame is the loss function (1.1), which we conclude does not provide a coherent basis for decision-theoretic set estimation. To substantiate this claim, we now present a class of loss functions that do not lead to this paradox.

3. Resolving the Loss Function Paradox

As we blame the undesirable behavior of $C_t'(\bar{x},s)$ on the loss function (1.1), we now attempt to resolve the paradox by investigating other loss functions. If a decision-theoretic set estimation theory is to be viable, we must find a loss function that both eliminates the paradox and is reasonable to an experimenter. To minimize complexity, we examine loss functions of the form (1.2), that is,

$$L_S(\theta,C) = S[\mathrm{vol}(C)] - I(\theta \in C)\,, \tag{3.1}$$

where $S(\cdot)$ is a continuous, increasing function. The class of losses (3.1) contains the linear loss (1.1) and, we will see that conditions on S can be derived to eliminate the paradoxical behavior.

We can first classify the unwanted behavior of the Bayes sets into the following three types:

1. The Bayes set is empty for some values of s.
2. The Bayes radius $st^*(s) \downarrow 0$ as $s \uparrow \infty$.
3. The Bayes radius $st^*(s)$ decreases for some range of s.

We focus on the generalized prior that leads to the t-interval, but it is clear that the same requirements are natural for other prior distributions. The three requirements are increasingly restrictive, but at least we should avoid the first paradox. A Bayes set from a proper prior should not be $\emptyset$ or $\Re$, other than in limiting cases. Even if there is no additional information from the sample, the prior alone should provide more than a "degenerate" interval.

As we will se in Section 4, prior distributions must be absolutely continuous with respect to Lebesgue measure on Θ, otherwise problems may arise. (In particular, there could

be problems even defining Bayes sets.) Under this assumption, the Bayes sets associated with (3.1) are HPD regions given by

$$C^{\pi}(\bar{x},s) = \{\theta: \pi(\theta \mid \bar{x},s) > k(\bar{x},s)\}.$$

This illustrates another advantage of working with losses of the form (3.1) rather than more general forms. More general losses may result in Bayes estimators that are not HPD regions, which can be considered undesirable.

First we show that volume and coverage probability must be weighted equally in order to avoid counterintuitive Bayes sets. Without loss of generality, we can assume that $S(0) = 0$, which makes the loss of $\emptyset$ equal to 0. In the following argument we will repeatedly use the fact that the posterior loss of a Bayes rule is nonpositive.

Proposition 3.1. *If* $S(+\infty) > 1$, *a paradox occurs in that the radius of the Bayes set approaches* 0 *as* $s \to +\infty$. *That is,*

$$\lim_{s \to +\infty} st^*(s) = 0 ,$$

where $t^*(s)$ *is the value of* t *that minimizes*

$$S(2ts/\sqrt{n}) - P(|T_{n-1}| \leq t) \tag{3.2}$$

and T_{n-1} *denotes a Student's* t *random variable with* n-1 *degrees of freedom.*

Proof. As $\lim_{v \to +\infty} S(v) = B > 1$, there exists v_0 such that $S(v) \geq 1$ for $v \geq v_0$. Therefore, necessarily,

$$2\frac{t^*(s)s}{\sqrt{n}} \leq v_0 ,$$

which implies $\lim_{s \to +\infty} t^*(s) = 0$. But then, the posterior probability satisfies

$$\lim_{s \to +\infty} P\left(|T_{n-1}| < t^*(s)\right) = 0.$$

Therefore, we must have $\lim_{s \to +\infty} \left(2\frac{t^*(s)s}{\sqrt{n}}\right) = 0$ which implies $\lim_{s \to +\infty} st^*(s) = 0$, proving the theorem. □

If $S(+\infty) < 1$, there is not the same undesirable behavior, but we choose to eliminate this case for the following reason. As $s \uparrow \infty$, eventually it will happen that

$$S[2t^*(s)s/\sqrt{n}] - P\left(|T_{n-1}| \leq t^*(s)\right) > S(+\infty) - 1, \tag{3.3}$$

which implies that the Bayes set will equal Θ for finite values of s. Although this behavior is not terrible, as we argued before, it is more desirable that the Bayes set not be the entire

space, except in the limit. Additionally, as the empty set and the entire parameter space are equally noninformative, they should receive the same weight.

Even the loss functions satisfying the condition

$$L_S(\theta, \emptyset) = L_S(\theta, \Theta) = 0 \tag{3.4}$$

may result in paradoxical behavior if the size function S grows too rapidly. This is illustrated in the following proposition.

Proposition 3.2. *If there exists* δ_0, ω_0 *such that*

$$S(\delta\omega) - P\left(|T_{n-1}| < \omega\right) \geq 0 \tag{3.5}$$

for $\delta \geq \delta_0$, $\omega \geq \omega_0$, *the solution of* (3.2) *satisfies*

$$\lim_{s \to +\infty} st^*(s) = 0 .$$

Proof. We have necessarily, for $2\frac{s}{\sqrt{n}} \geq \delta_0$, $t^*(s) \leq \omega_0$. But then, if $\lim_{s \to +\infty} t^*(s) \neq 0$, it follows that

$$\lim_{s \to +\infty} S\left(2\frac{st^*(s)}{\sqrt{n}}\right) - P\left(|T_{n-1}| < t^*(s)\right) = 1 - \lim_{s \to +\infty} P\left(|T_{n-1}| < t^*(s)\right),$$

which is strictly positive (as $t^*(s) < \omega_0$). This contradicts the Bayes assumption, and implies $\lim_{s \to +\infty} t^*(s) = 0$. A similar argument will establish $\lim_{s \to +\infty} st^*(s) = 0$. □

An example of a loss function that satisfies both conditions (3.4) and (3.5), hence displays undesirable behavior, is given below.

Example 3.1. Consider the size function

$$S_a(v) = 1 - e^{-av^2/2},$$

which results in a loss function of the form (3.1) satisfying (3.4) and (3.5). The derivative of the posterior expected loss, with respect to t is

$$2\frac{as^2 t}{n} e^{-as^2t^2/n} - 2f_{n-1}(t) .$$

This expression is negative for t close to 0 and t large, so the solution of (3.2) is the smallest solution of

$$\frac{as^2 t}{n} e^{-as^2t^2/n} = f_{n-1}(t) . \tag{3.6}$$

As $s \to +\infty$, the smallest solution of (3.6) is going to 0. It can then be established, as in the previous propositions, that $\lim_{s \to +\infty} st^*(s) = 0$. □

For the t-interval, we can exhibit sufficient conditions for a loss function to be non-paradoxical. Under the assumption that $S(\cdot)$ is continuously differentiable, we have the following result.

Proposition 3.3. *Let the size function of the loss* (3.1) *satisfy* (3.4) *and be continuously differentiable. If either*

$$\frac{\partial}{\partial t}\left\{S(2ts/\sqrt{n}) - P\left(|T_{n-1}| \leq t\right)\right\} < 0 \quad \textit{for sufficiently small } t > 0,$$

or

$$\frac{\partial}{\partial t}\left\{S(2ts/\sqrt{n}) - P\left(|T_{n-1}| \leq t\right)\right\} > 0 \textit{ for } t > M\,,$$

where M is a constant, then $t^*(s)$ is a solution to

$$S'[2t^*(s)s/\sqrt{n}]\tfrac{s}{\sqrt{n}} = f_{n-1}[t^*(s)].$$ □

We now have, for a class of size functions, a characterization of necessary and sufficient conditions on the loss function to eliminate paradoxical behavior. The next example gives a particularly simple, and appealing, size function that satisfies these conditions.

Example 3.2. Consider a size measure of the form

$$S_a(v) = \frac{v}{a + v}. \tag{3.7}$$

Notice that this size measure decreases as the Cauchy distribution (Student's t with one degree of freedom) and thus is well suited for use with t densities. This is because, as we have seen, we would like the size function to decrease more slowly than the density function.

Figure 3.1 about here

The derivative of the posterior expected loss is

$$\frac{2as/\sqrt{n}}{\left(a + 2st/\sqrt{n}\right)^2} - 2f_{n-1}(t)\,, \tag{3.8}$$

which has either one or two zeros (in t), as shown in the proof of Theorem 4.3. If there is one zero, it corresponds to the minimum. If there are two zeros, the larger zero is the minimum. In both cases, it is easy to see that $t^*(s)$ is going to $+\infty$ with s. Thus,

$\lim_{s \to +\infty} st^*(s) = +\infty$, and there is no paradoxical behavior for this size function. Moreover, the third requirement for non-paradoxical behavior, that $st^*(s) \uparrow s$ is also satisfied. In Figure 3.1a and Figure 3.1b we illustrate the behavior of $t^*(s)$ and $st^*(s)$ for different values of a. From Figure 3.1 we see that $t^*(s)$ is not a monotone function of s but Figure 3.2 shows that the more important quantity, the Bayes radius $st^*(s)$, is a monotone function. □

4. Some Decision-Theoretic Consequences

With an acceptable loss function for set estimation, one that takes values in the real numbers, we could now investigate some standard decision-theoretic properties such as minimaxity or admissibility. This, as mentioned before, was one of the goals of searching for a loss function suitable for set estimation. In this section we present two consequences of such decision-theoretic investigations. A more thorough development of decision-theoretic properties may be found in Casella, Hwang and Robert (1990).

There are some technical difficulties associated with the comparisons of set estimators using a loss function, as mentioned by Joshi (1969). For a set estimator C of θ, the new set estimator $C' = C \cup \{\theta_0\}$ dominates C in risk. Hence, according to the usual definition, there can be no admissible estimators. To avoid these difficulties, we only consider *Lebesgue-admissiblity.* That is, the set C is Lebesgue-admissible if, for any set C', $R(\theta,C) - R(\theta,C') \geq 0$ (a.e.) $\Rightarrow R(\theta,C) - R(\theta,C') = 0$ (a.e). This is why we only consider priors that are absolutely continuous with respect to Lebesgue measure.

With respect to Lebesgue-admissibility, all of the regular decision theory results hold. For example, the admissible procedures form a minimal complete class, Bayes procedures are HPD regions and, if unique, are admissible. Also, there exists a minimax rule, and suitable limits of Bayes procedures are a complete class. Some results of this type can be found in Brown (1977), and others are in Casella, Hwang and Robert (1990).

We focus here on two decision theoretic properties of the sets arising from the size function (3.7). We first give necessary and sufficient condition for a Bayes set to be nontrivial (a Bayes set is *nontrivial* if it is neither $\emptyset$ nor Θ with positive posterior probability). In Section 4.2 we exhibit a minimum coverage probability of the Bayes t-interval sets.

4.1. Nontrivial Bayes Sets

Recall from Section 2 that we required reasonable loss functions, when used with proper prior distributions, to produce Bayes sets that were nontrivial. For the size function (3.7), $S_a(v) = v/(v+a)$, which has already been seen to be reasonable, we are able to provide a simple sufficient condition for this to hold.

Theorem 4.1. *If* $\{\theta: \pi(\theta|x) > (1/a)\}$ *has positive Lebesgue measure, the Bayes rule* $C_x^\pi = \{\theta: \pi(\theta|x) > k\}$ *against the loss function* $L(\theta, C) = S_a(vol(C)) - I(\theta \in C)$ *is nontrivial, since* $k = k(x) < 1/a$.

Proof. The derivative of the posterior expected loss of a set $C_X^\pi(k) = \{\theta: \pi(\theta|x) > k\}$ is

$$\frac{\partial}{\partial k} L(C_X^\pi(k)|x) = \left(k - S_a'[\mathrm{vol}(C_X^\pi(k))]\right) \int_{\{\pi(\theta|x)=k\}} |\nabla \pi(\theta|x)|^{-1}\, ds$$

$$= \left(k - \frac{a}{\left(a + \mathrm{vol}(C_X^\pi(k))\right)^2}\right) \int_{\{\pi(\theta|x)=k\}} |\nabla \pi(\theta|x)|^{-1} ds,$$

where ds represents the infinitesimal surface area of the set $\{\theta: \pi(\theta|x) = k\}$, and $\nabla\pi(\theta|x)$ is the gradient of $\pi(\theta|x)$ for fixed x. Since

$$\frac{a}{\left(a + \mathrm{vol}(C_X^\pi(k))\right)^2} \le \frac{1}{a} \quad \text{for every } k,$$

if $\int_{\{\pi(\theta|x)=k\}} |\nabla\pi(\theta|x)|^{-1} ds$ is different from 0 for some $k \ge \frac{1}{a}$, $L(C_X^\pi(k)|x)$ will be increasing for $k \ge 1/a$ and the minimum, in k, of the posterior loss will occur for $k < \frac{1}{a}$. □

Note that for any value of a and any sample distribution, there are always prior distributions satisfying $\pi(\theta|x) < 1/a$. However, this does not necessarily imply that the associated Bayes set is empty. The situation can be even more complicated, as the next corollary shows.

Corollary 4.2. *If* $\{\theta: \pi(\theta|x) > k_0\}$ *has Lebesgue measure 0 for* $k_0 < \frac{1}{a}$, *the posterior loss* $L(C_x^\pi(k)|x) = S_a[vol(C_x^\pi(k))] - P(\theta \in C_x^\pi(k)|x)$ *is decreasing as* $k \uparrow k_0$.

Proof. As $\mathrm{vol}(C_X^\pi(k))$ is a continuous function of k, for every $\epsilon > 0$, there exists $\delta > 0$ such that $|k-k_0| < \delta$ implies

$$\left| \frac{a}{\left(a + \mathrm{vol}(C_X^\pi(k))\right)^2} - \frac{1}{a} \right| < \epsilon .$$

Therefore, if $k_0 - k < \delta$,

$$k - S'_a[\mathrm{vol}(C_X^\pi(k))] = k - \frac{a}{\left(a + \mathrm{vol}(C_X^\pi(k))\right)^2} < k - \frac{1}{a} + \epsilon < 0 .$$

□

Corollary 4.2 does not necessarily imply that $\emptyset$ is the Bayes rule in this case, as it is still possible that $L(C_x^{\pi}(k)|x)$ is also decreasing for k close to 0 (see Example 4.1).

Example 4.1. Suppose $X \sim N_p(\theta, I)$ and $\theta \sim N_p(0, \tau^2 I)$. Then $\pi(\theta|x)$ is $N_p\left(\eta x, \eta I\right)$ and the Bayes set is $C_x^{\pi} = \left\{\theta: |\theta - \eta x^2| \le c\right\}$ where $\eta = \tau^2/(\tau^2+1)$. If a satisfies $a > (2\pi)^{p/2} \ge (2\pi\eta)^{p/2}$, the Bayes set is never empty, according to Theorem 4.1. If $a < (2\pi\eta)^{p/2}$ then $\left\{\pi(\theta|x) > (1/a)\right\}$ is empty; however, there still exists a nontrivial Bayes set (see Casella, Hwang and Robert, 1990). □

4.2. The Range of Bayes Sets.

In Example 3.2 we saw that the size function (3.7) results in non-paradoxical behavior for the Bayes t-interval. In addition, here we investigate the range of the values of the posterior coverage probabilities as a function of a. This range is of interest in evaluating the flexibility of the loss function in answering the needs of the experimenter. We want the probability to have a reasonable range, but we would also like to give the assurance of a lower bound.

Theorem 4.3. *For the size function* (3.7), $S_a(v) = v/(v+a)$, *the minimum coverage probability for the Bayes t-interval is* $\frac{1}{2}$, *regardless of the value of* a.

Proof. The Bayes set is

$$C_{t^*}(\bar{x},s) = \left\{\mu: |\bar{x} - \mu| < t^* \frac{s}{\sqrt{n}}\right\}$$

where t^* is the solution of

$$\min_t \left\{\frac{2ts/\sqrt{n}}{a + 2ts/\sqrt{n}} - P\left(|T_{n-1}| < t\right)\right\}. \tag{4.1}$$

Differentiation of the expression in the braces shows that t^* is a solution of

$$\frac{a2s/\sqrt{n}}{\left(a + 2ts/\sqrt{n}\right)^2} - 2\frac{\Gamma(n/2)}{\left((n-1)\pi\right)^{1/2}\Gamma(n-1/2)}\left(1 + \frac{t^2}{n-1}\right)^{-n/2} = 0. \tag{4.2}$$

Now, some algebra will show

$$\text{sign}\left[\frac{\partial}{\partial t}\left\{\frac{\left(1 + \frac{t^2}{n-1}\right)^{n/2}}{\left(a + 2ts/\sqrt{n}\right)^2}\right\}\right] = \text{sign}\left[2s\frac{n-2}{n-1}\frac{t^2}{\sqrt{n}} + \frac{n}{n-1}at - 4\frac{s}{\sqrt{n}}\right],$$

which implies that equation (4.2) has one solution if $s < a\sqrt{n}\Gamma(n/2)/\sqrt{(n-1)\pi}\,\Gamma((n-1)/2)$ and two otherwise. In this latter case, t^* is the larger value.

Furthermore, it is straightforward to check that

$$\operatorname{sign}\left[\frac{\partial}{\partial s}\, t^*(s)\right] = \operatorname{sign}\left[2\frac{2t^*(s)\frac{s}{\sqrt{n}}}{a + 2t^*(s)s/\sqrt{n}} - 1\right],$$

and therefore $t^*(s)$ is decreasing if $(2t^*(s)s/\sqrt{n})/(a + 2t^*(s)s/\sqrt{n}) < 1/2$ and increasing otherwise. Thus, it follows that the unique minimum value satisfies

$$\min_s \frac{2t^*(s)s/\sqrt{n}}{a + 2t^*(s)s/\sqrt{n}} = \frac{1}{2}.$$

Remembering that the posterior expected loss (4.1) must be negative, we obtain

$$\min_s P\left(|T_{n-1}| < t^*(s)\right) \geq \frac{1}{2}. \tag{4.3}$$

Thus, the minimum coverage probability (either frequentist or Bayesian) must be at least 1/2. □

5. Conclusions

Our interest in decision-theoretic set estimation stemmed from Jim Berger's Student's *t* paradox of Section 2. The fact that such a time-honored procedure could be dominated by an obviously silly procedure convinced us that the decision-theoretic approach to set estimation needed a long look taken at it. We have learned that a nonlinear size function not only can eliminate the paradox, but also such size functions are, in general, analytically tractable.

There is an important relationship between the size function and the underlying sample density. The fact that the size function of (3.7), which has a Cauchy-like tail, is a reasonable one for the *t* distribution is partially due to the fact that the rate of change at the tails is larger than that of the Student's *t* cdf. (Or, that the derivative of S decreases more slowly that Student's *t* pdf.) Also, although we have focused on Student's *t*, the requirements we introduced apply to every distribution. In particular, size functions $S(\cdot)$ should always satisfy

$$S(0) = 0 \ , \ S(+\infty) = 1.$$

The choice of loss function, that is, the manner in which size and coverage probability are to be combined, is of major importance. We have no overwhelming reason to prefer a loss function using (3.7) except for its simplicity and performances in the cases considered. Other loss functions, some of which are particularly applicable to bounded parameter spaces, are

discussed in Casella, Hwang and Robert (1990).

To approach the entire set estimation problem decision-theoreticall, instead of being concerned with separate measures of size and coverage, leads to combining these measures in a single-valued loss function. In this decision-theoretic setting, after specifying the loss function, the model and the data specify the size and coverage. A disadvantage of this approach is that it requires more careful thinking about the relative importance of these measures, while an advantage is that it allows interplay between size and coverage probability.. Also, we can easily define typical decision-theoretic quantities like admissibility and minimaxity, definitions which are ambiguous with a vector-valued loss, for example.

The decision-theoretic approach to set estimation provides us with powerful methods, letting us appropriately balance size and coverage. Thus far, three main forms have been examined; the vector loss approach, the linear combination loss as in (1.1), and the nonlinear combination loss examined here. Although the first two approaches can sometimes yield reasonable answers, they have disadvantages. We believe that the use of a nonlinear size function provides the most attractive alternative. The nonlinear size function provides coherent behavior of optimal set estimators, provides nontrivial Bayes sets, and can give minimum coverage guarantees.

Acknowledgement. We thank Larry Brown, Arthur Cohen and Colin Blyth for their valuable comments. Also, the suggestions of the referees and associate editor helped this paper to evolve from an earlier draft.

References

Berger, J. O. (1980). A Robust Generalized Bayes Estimator and Confidence Region for a Multivariate Normal Mean. *Ann. Statist.* 8, 716-761.

Blyth, C.R. (1951). On Minimax Statistical Decision Procedures and Their Admissibility. *Ann. Math. Statist.* 22, 22-42.

Blyth, C.R. and Hutchinson, D. W. (1961). Tables of Neyman-Shortest Confidence Intervals for the Binomial Parameter. *Biometrika* 47, 381-391.

Brown, L.D. (1966). On the Admissibility of One or More Location Parameters. *Ann. Math. Statist.* 37, 1087-1136.

Brown, L.D. (1977). Closure Theorems for Sequential-Design Processes. In *Statistical Decision Theory and Related Topics II* (S.S. Gupta and D.S. Moore, eds). Academic Press, 57-91.

Casella, G. and Hwang, J.T. (1982). Evaluating Confidence Sets using Loss Functions. Biometrics Unit Technical Report BU-773-M, Cornell University. To appear in *Statistica Sinica.*

Casella, G. and Hwang, J.T. (1983). Empirical Bayes Confidence Sets for the Mean of a Multivariate Distribution. *J. Amer. Statist. Assoc.* 78, 688-698.

Casella, G. and Hwang, J.T. (1986). Confidence Sets and the Stein Effect. *Comm. Statist. Theor. Meth. Special Issue on Stein Estimation*, 15(7) 2043-2063.

Casella, G., Hwang, J.T., and Robert, C. (1990). Loss Functions for Set Estimation. Biometrics Unit Technical Report BU-999-M, Cornell University.

Cohen, A. and Strawderman, W.E. (1973a). Admissibility Implications for Different Criteria in Confidence Estimation. *Ann. Statist* 1, 363-366.

Cohen, A. and Strawderman, W.E. (1973b). Admissible Confidence Interval and Point Estimation for Translation or Scale Parameters. *Ann. Statist* 1, 545-550.

Cohen, A. and Sackrowitz, H.B. (1984). Decision Theory Results for Vector Risks with Applications. *Statistics and Decisions*, Supplement Issue No. 1, 159-176.

Hooper, P. (1982). Invariant Confidence Sets with Smallest Expected Measure. *Ann. Statist.*

10, 1283-1294.

Hwang, J.T. and Casella, G. (1982). Minimax Confidence Sets for the Mean of a Multivariate Normal Distribution. *Ann. Statist.* 10, 868-881.

Hwang, J.T. and Casella, G. (1984). Improved Set Estimators for a Multivariate Normal Mean. *Statistics and Decisions, Supplement Issue* 1, 3-16.

Joshi, V.M. (1967). Admissibility of the Usual Confidence Set for the Mean of a Multivariate Normal Population. *Ann. Math. Statist.* 38, 1868-1875.

Joshi, V.M. (1969). Admissibility of the Usual Confidence Set for the Mean of a Univariate or Bivariate Normal Population. *Ann. Math. Statist.* 40, 1042-1067.

Meeden, G. and Vardeman, S. (1985). Bayes and Admissible Set Estimation. *J. Amer. Statist. Assoc.* 80, 465-471.

Neyman, J. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Phil. Trans., A*, 236, 330-380.

Pratt, J.W. (1961). Length of Confidence Intervals. *J. Amer. Statist. Assoc.* 56, 549.

Stein, C. (1962). Confidence Sets for the Mean of a Multivariate Normal Distribution (with discussion). *J. Royal. Statist. Soc.* B. 24, 573-601.

Sterne, T.E. (1954). Some Remarks on Confidence or Fiducial Limits. *Biometrika* 41, 275-278.

Tate, R.F. and Klett, G.W. (1959). Optimal Confidence Intervals for the Variance of a Normal Distribution. *J. Amer. Statist. Assoc.* 54, 674-682.

Wilks, S.S. (1938). Shortest Average Confidence Intervals from Large Samples. *Ann. Math. Statist.* 9, 272.

Winkler, R. L. (1972). A Decision-Theoretic Approach to Interval Estimation. *J. Amer. Statist. Assoc.* 67, 187-191.

Statistics Center
Caldwell Hall
Cornell University
Ithaca, NY 14853

L.S.T.A - Tour 45-55
Université Paris VI
4, Place Jussieu
F-75252 Paris Cedex 05

Figure 3.1a. For the Bayes set against the loss (3.1) with S(v) = v/(a+v), graphs of $t^*(s)$ for $a = 1/2$ (solid), 2 (long dashes), 5 (dotted), 20 (short dashes). The distribution is Student's t with 25 degrees of freedom

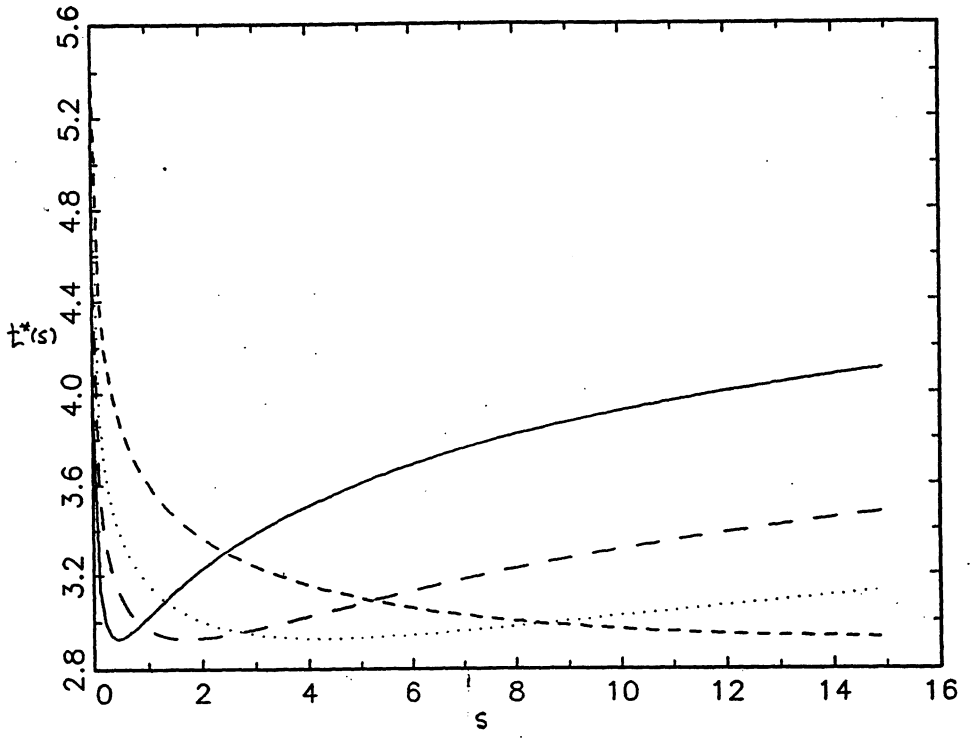

Figure 3.1b. For the Bayes set against the loss (3.1) with S(v) = v/(a+v), graphs of $st^*(s)$ for $a = 1/2$ (solid), 2 (long dashes), 5 (dotted), 20 (short dashes). The distribution is Student's t with 25 degrees of freedom